

Customer Churn Prediction

Problem

Telecom companies lose revenue when customers cancel their subscriptions. The goal of this project is to predict which customers are likely to churn and evaluate different targeting strategies to reduce financial loss.

Dataset

Telco Customer Churn dataset (~7,000 customers).

Target:

- 1 → Churn
- 0 → No churn

Churn rate: ~26.6%

EDA Summary

Some clear patterns observed:

- Customers with month-to-month contracts churn much more frequently.
 - Low tenure customers are more likely to churn.
 - Higher monthly charges increase churn risk.
 - Fiber optic users show higher churn rates.
 - Customers without tech support are more likely to churn.
 - Gender has almost no effect.
-

Feature Engineering

Added features:

- $\text{AvgChargePerMonth} = \text{TotalCharges} / \text{tenure}$
- HasInternet (binary)
- SupportScore (number of enabled support services)

Categorical features were one-hot encoded. Train/test split was done with stratification (80/20).

Final dataset: 32 features.

Models

Trained:

- Logistic Regression

- Random Forest
- XGBoost

Evaluation metrics:

- ROC-AUC
- PR-AUC

XGBoost performed best overall.

Threshold Optimization

Default threshold (0.5) produced low churn recall.

After tuning for F1 score:

- Best threshold: 0.24
- F1 score: 0.596

Lowering the threshold improved churn detection significantly.

Cost-Based Evaluation

Assumptions:

- Losing a customer costs \$300
- Campaign cost per customer is \$50

Expected cost:

- Threshold 0.5 → \$73,000
- Threshold 0.24 → \$56,450

Threshold optimization reduced expected loss by \$16,550.

Top-K Targeting

Simulated a realistic constraint: Only top 300 highest-risk customers are targeted.

Results:

- Precision@300: 59%
- Recall@300: 47%
- Base churn rate: 26.6%

The model selects customers more than 2x better than random selection.

Model Interpretation

SHAP analysis shows the main churn drivers:

- Low tenure
- Month-to-month contracts
- High monthly charges
- Fiber optic service
- Electronic check payment method
- Lack of tech support

The model aligns well with observed patterns in the data.

Results

Model comparison showed that XGBoost performed best in terms of ROC-AUC and PR-AUC.

Key results:

- Best decision threshold: 0.24
- F1 score: 0.596
- Expected cost at threshold 0.5: \$73,000
- Expected cost at threshold 0.24: \$56,450
- Estimated cost reduction: \$16,550

Top-300 targeting simulation:

- Precision@300: 59%
- Recall@300: 47%
- Base churn rate: 26.6%

This means that targeting only 300 high-risk customers captures nearly half of all churn cases and performs more than 2x better than random selection.

Summary

This project covers:

- Data exploration
- Feature engineering
- Model comparison
- Threshold tuning
- Cost-based analysis
- Budget-constrained targeting
- Model interpretation

The final model provides both predictive performance and practical business insights.